

Noise-aware fine-tuning as a hardware noise margin: Cross-layer co-design for robust low bit RRAM in-memory computing toward space-borne edge inference

Yang Tian¹, Dunan Hu², Lingzi Zhang¹, Lili Gao¹, Kequan Xia³, Tuo Shi⁴, Jianguo Lu²

¹ Zhejiang Laboratory, Hangzhou 311121, China, ² State Key Laboratory of Silicon and Advanced Semiconductor Materials, School of Materials Science and Engineering, Zhejiang University, Hangzhou 310058, China, ³ Department of Mechanical Engineering, Imperial College London, SW7 2AZ, London, United Kingdom, ⁴ College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310058, China

Corresponding author: Tuo Shi, Jianguo Lu, shituo@zhejianglab.com, lujianguo@zju.edu.cn

Resistive Random Access Memory (RRAM)-based Computing in-memory (CIM) accelerators promise orders of magnitude energy savings for space-borne edge inference. At the low bit precision imposed by tight power budgets, however, device-level conductance noise competes directly with the quantization signal margin. While conventional Quantization-Aware Training (QAT) can adapt neural networks to low bit hardware, it remains insufficient against physical conductance fluctuations, often leading to severe accuracy degradation in regression tasks. We therefore introduce noise-aware fine-tuning as a cross-layer robustness primitive, a three-stage pipeline (PQAT-FT) in which sparse regularization and noise injection reshape the weight distribution to carve unused conductance headroom, interrupting layer-by-layer error propagation. Measured on a fabricated 8K RRAM CIM test chip over five programming rounds (mean $\pm$ SD), PQAT-FT consistently reduces the 4-bit inference error across both bounded and unbounded activations, lowering the RMSE for Tanh, Sigmoid, Leaky ReLU, and ReLU by 64.2%, 18.7%, 25.3%, and 36.8%, respectively. Notably, Tanh exhibits the largest relative improvement over its QAT baseline. These findings establish noise-aware fine-tuning as an effective primitive for low bit RRAM CIM across the four activations and bit widths tested on this hardware platform, toward future space-borne edge systems.

Keywords: Computing in-memory, cross-layer co-design, noise-aware fine-tuning, resistive random access memory, quantization-aware training, space edge AI

1. INTRODUCTION

Onboard artificial intelligence for space-borne edge computing must maintain high inference fidelity under strict power budgets and severe radiation [1, 2, 3]. Monthly sunspot number prediction is a critical space-weather forecasting task [4]. The prediction results directly determine the timing of radiation protection for satellite electronics and astronauts. Models must execute continuous and autonomous inference on orbit. This

is because the ground-station telemetry round-trip delay usually exceeds the 15–30 minute golden warning window required for protective actions. Cosmic radiation also induces volatile Single Event Upsets (SEUs) in flight [5, 6]. Analog Computing in-memory (CIM) accelerators based on Resistive Random Access Memory (RRAM) offer a practical path to meet these energy and latency constraints. RRAM CIM performs matrix-vector multiplications directly inside the crossbar array via analog current summation. This mechanism eliminates the von Neumann data movement bottleneck, yielding clear area and energy efficiency gains over digital Complementary Metal–Oxide–Semiconductor (CMOS) processors [7, 8]. Beyond energy efficiency, non-volatile RRAM cells are attractive for space deployment because previous studies report radiation tolerance in HfO_2 -based RRAM devices: heavy-ion irradiation experiments report no single event upsets in the stored conductance states under realistic LET exposures [9], making the crossbar a candidate for radiation-robust on orbit weight storage.

Digital processors can represent synaptic weights precisely via high precision floating-point values. In contrast, constrained by inherent conductance fluctuations and device non-idealities, physical RRAM cells can only maintain a limited number of stable discrete conductance states, making it difficult to accurately represent full-precision floating-point weights [10, 11]. Advanced Multi-Level Cell (MLC) technologies improve resolution, but the number of stable conductance levels remains strictly limited compared to the digital domain. Chip-level experimental demonstrations of memristor-based feature learning [12] and neuro-fuzzy systems in crossbar arrays [13] have further validated the physical realizability of in situ intelligent computing on RRAM hardware. Under such constrained physical signal-to-noise ratios, the quantized conductance state spacing competes directly with intrinsic device non-idealities. These non-idealities include Random Telegraph Noise (RTN) that fluctuates across conductance states, programming errors, conductance drift, and I-V nonlinearity [14, 15]. Consequently, structural design choices traditionally optimized for unconstrained digital domains begin to dictate on-chip inference accuracy.

To close the gap between simulation and silicon, existing RRAM CIM research has pursued co-optimization along two main lines: improving memristive-material uniformity with fault-aware mapping and peripheral compensation, and developing Hardware-Aware Training (HAT) that injects calibrated noise during Quantization-Aware Training (QAT) [16]. Probabilistic and error-aware training strategies further tolerate memristive variability [17], and prior work [18] has shown the viability of integrated hardware. These routines, however, share an implicit limitation: they primarily focus on mapping accuracy and noise emulation, while overlooking how the optimization trajectory shapes the weight distribution

within the physical conductance window. Conventional quantization routines often disperse weights across the full operational range. Without intentional margins, unavoidable device fluctuations easily drive weights into degenerate regimes, triggering multiplicative error accumulation through downstream matrix multiplications [19, 20, 21]. Crucially, on-chip robustness is governed not merely by architectural choice but fundamentally by whether the training strategy proactively carves vacant conductance headroom to absorb physical perturbations.

We therefore propose noise-aware fine-tuning as a hardware noise margin. Fig. 1 outlines the proposed cross-layer co-design: our three-stage pipeline, Pretrained Quantization-Aware Training followed by noise-aware Fine-Tuning (PQAT-FT), reshapes the weight landscape through sparse regularization and noise injection. During inference, this carved conductance headroom functions as an intrinsic buffer that absorbs device-level fluctuations before they can propagate across layers. We demonstrate on a fabricated 8K RRAM CIM test chip that PQAT-FT consistently lowers 4-bit inference error across diverse network activations, with bounded functions like Tanh serving as an optimal complement that unlocks the largest relative improvement. Furthermore, under this training constraint, we reveal a non-monotonic dependence on weight quantization precision under the tested conditions, where a coarser 4-bit grid can surpass 6-bit precision on physical hardware. We validate the framework on a multi-cycle solar sunspot dataset (1750–2019) and further probe its boundary on a 2D space-borne remote-sensing benchmark (EuroSAT) under simulated noise, providing cross-task evidence for the proposed methodology.

2. METHODS

2.1 RRAM CIM hardware platform

The test chip integrates an 8K RRAM crossbar array in a $\text{TiN}/\text{TaOx}/\text{HfOx}/\text{TiN}$ (TaOx/HfOx bilayer RRAM) process. Each cell stores one analog neural network weight as a programmed conductance (one weight per cell, regardless of the quantization bit precision). The programmed conductances span an operational window from $G_{\min} = 50 \mu\text{S}$ to $G_{\max} = 250 \mu\text{S}$. Peripheral circuits provide 13-bit DAC input encoding and 16-bit ADC readout. Prior to mapping, the trained weights of each layer are normalized by the layer-wise maximum absolute magnitude, so that raw (quantized) weights that may exceed unity are rescaled into the mapped weight $w \in [-1, 1]$ (the raw weight distributions are shown in Fig. 2c and Fig. 2d). Each such mapped weight w is then written to a single device conductance in a reference-column scheme, $G = G_{\text{base}} + \Delta G \cdot w$, with

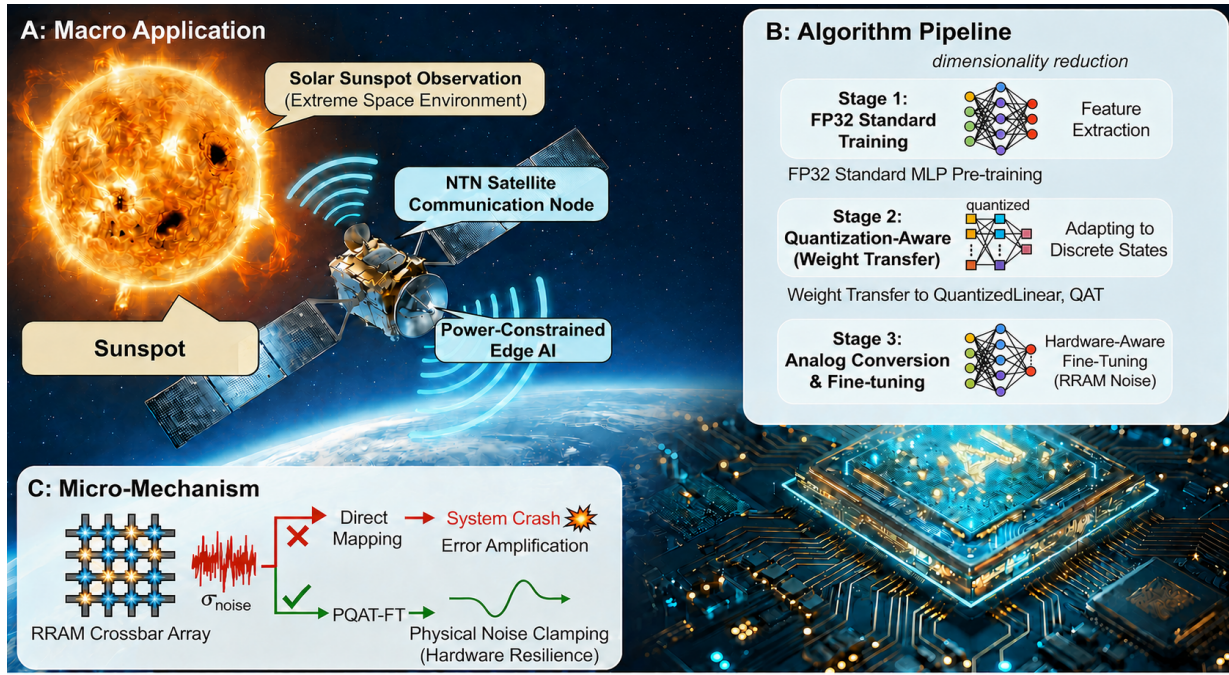


Figure 1 – Cross-layer co-design framework for robust low bit RRAM CIM. (a) Space-borne edge AI motivates RRAM CIM deployment under strict energy constraints. (b) A conventional quantized pipeline amplifies device-level conductance noise through unbounded error propagation. (c) The proposed PQAT-FT pipeline reshapes weights to leave a noise margin and suppresses error accumulation, validated by on-chip measurement. (Illustration created with the assistance of Google Gemini and ChatGPT; content verified by the authors.)

base conductance $G_{\text{base}} = (G_{\text{min}} + G_{\text{max}})/2 = 150 \mu\text{S}$ and conductance swing $\Delta G = (G_{\text{max}} - G_{\text{min}})/2 = 100 \mu\text{S}$, so that $w = 0$ maps to $150 \mu\text{S}$ and the end-points $|w| = 1$ map to the window edges; a reference column supplies the base current $I_{\text{base}} = G_{\text{base}} \sum_i x_i$ that is subtracted from the array column current so that signed (positive/negative) weights are represented. On-chip inference encodes input vectors as voltage pulses, sums currents via Kirchhoff's law, and digitizes column outputs. A single analog read operation applies a 0.185 V , 20 ns pulse to each selected cell. The cell-level electrical energy of this operation is $E_{\text{cell}} = t_{\text{pulse}} V_{\text{read}}^2 G_{\text{base}} = 20 \text{ ns} \times (0.185 \text{ V})^2 \times 150 \mu\text{S} = 103 \text{ fJ}$, which is the array electrical contribution only and excludes peripheral and system-level overhead; the peripheral-inclusive per-operation energy (1.65 pJ/op) is reported separately in the efficiency comparison. The cell-level electrical contribution follows the same per-operation methodology used on the identical TiN/TaOx/HfOx/TiN platform [12].

2.2 Network architecture and quantization

The predictor is a three-layer MLP ($7 \rightarrow 20 \rightarrow 20 \rightarrow 1$) with dropout (rate 0.2). Seven lagged sunspot observations constructed by phase-space reconstruction with a delay of 37 months are mapped to the subsequent month's value. The same activation function is applied at every layer, including the output. For ReLU, the final output is $\max(0, z)$, negative pre-activations are rectified

to zero, and positive pre-activations pass through unattenuated. For Tanh, output values are constrained to $[-1, 1]$ before inverse normalization to physical sunspot units.

Weight quantization uses per-tensor uniform quantization with Straight-Through Estimator (STE):

$$s = \frac{\max(\mathbf{W}) - \min(\mathbf{W})}{2^b - 1} \quad (1)$$

$$z = \text{clamp}(\text{round}(-\min(\mathbf{W})/s), 0, 2^b - 1) \quad (2)$$

$$\hat{\mathbf{W}} = s \cdot [\text{clamp}(\text{round}(\mathbf{W}/s + z), 0, 2^b - 1) - z] \quad (3)$$

The forward pass uses the dequantized weight $\hat{\mathbf{W}}$; the backward pass propagates gradients through the rounding operator unchanged. Only weights are quantized.

2.3 PQAT-FT: Three-stage training pipeline

The proposed pipeline (Fig. 1b) consists of three sequential stages designed to jointly optimize weight quantization and activation boundaries under hardware noise.

Stage 1 – FP32 Pretraining. The model is pretrained for 300 epochs in full 32-bit floating-point precision. Optimization is performed using AdamW with an initial learning rate of 0.001 and a weight decay of 10^{-4} . The learning rate schedule follows cosine annealing with

a minimum learning rate of $\eta_{\min} = 10^{-6}$. The training objective employs Mean Squared Error (MSE) loss.

Stage 2 – Quantization-Aware Training (QAT). The well-trained FP32 weights are transferred to the quantized network for subsequent quantization-aware training over 200 epochs at a learning rate of 0.0005. The Straight-Through Estimator (STE) is utilized to enable gradient backpropagation across quantization operations, which reorganizes the weight distribution to maintain inference fidelity under low bit quantization.

Stage 3 – Simulated noise-aware fine-tuning. The quantized model obtained from Stage 2 is imported into the IBM aihwkit platform (v0.9.2) [22] to construct a behavioral analog RRAM simulation model. Network fine-tuning is conducted for another 200 epochs using the AdamW optimizer at a learning rate of 0.0005, with L1 sparse regularization ($\lambda = 10^{-5}$, excluding bias terms) to encourage compact conductance distributions suitable for RRAM mapping. During fine-tuning, additive zero-mean Gaussian noise with $\sigma_n = 0.1$ (state-independent) is injected into each synaptic weight at every forward pass. This injection serves purely as a training time robustness regularizer that encourages the optimizer to leave margin in the weight distribution; it is not a model of the space radiation environment. At inference, evaluation uses the aihwkit device-level noise model, state-independent programming noise with a conductance-dependent standard deviation, rather than the additive Gaussian, to reflect realistic analog deployment. After noise-aware fine-tuning, the model is restored to a digital format and directly deployed on the physical RRAM Compute-In-Memory (CIM) chip without additional on-chip retraining. The optimal model checkpoint is determined by the best validation loss.

2.4 Data protocol and evaluation

The dataset is the monthly sunspot number series spanning 1750-01 to 2019-12 (3 240 months). Each sample is formed by phase-space reconstruction with a lag of 37 and an embedding dimension of 7, mapping seven lagged observations to the subsequent month, yielding 3 017 supervised samples. The series is split strictly in time, without shuffling: the first 2 753 samples form the training set, the following 132 months (1998-01 to 2008-12) form the validation set used only for early stopping, and the final 132 months (2009-01 to 2019-12) form the held-out test set, which is never seen during training. The entire series is normalized to $[0, 1]$ by a Min–Max scaler fitted exclusively on the training segment to prevent information leakage, and all reported errors are computed in the original physical units by inverse transformation. For each trained model, the reported RMSE is aggregated over multiple Monte-Carlo programming

trials (simulation) or repeated physical programming rounds (hardware), whose statistics are defined in the corresponding table captions. The five hardware rounds reflect repeated programming/inference variability on the same chip, not chip-to-chip variation.

The 2D benchmark is the EuroSAT land-use classification dataset (27 000 Sentinel-2 RGB patches, ten classes). Each image is resized to $64 \times 64 \times 3$ and flattened to a 12 288-dimensional input vector. The dataset is split 70/15/15 into training (18 900), validation (4 050), and test (4 050) sets in a stratified manner with a fixed random seed, and the inputs are standardized by a Z-score transform whose mean and variance are estimated on the training split only, so that no test statistics leak into preprocessing. The classifier is a three-layer MLP ($12288 \rightarrow 256 \rightarrow 128 \rightarrow 10$) with batch normalization and dropout (rate 0.2), matching the depth of the 1D predictor; Leaky ReLU is used deliberately, as an unbounded activation, to test whether the benefit extends beyond bounded functions on a vision task. Random flips and 90° rotations are applied during training.

The same three-stage protocol is followed, with two cost-driven deviations. The analog noise-aware fine-tuning stage is run for 60 epochs over a fixed random 5 000-sample subset drawn per epoch, because the EuroSAT training set is substantially larger than the sunspot training set and full analog training is computationally prohibitive; the pretraining stage additionally uses label smoothing of 0.1. All other settings, per-tensor uniform weight quantization at 4 bits, L1 sparse regularization ($\lambda = 10^{-5}$), AdamW with cosine annealing, and the injected fine-tuning noise $\sigma_n = 0.1$, are identical to the 1D configuration. Inference accuracy on the test split is evaluated with the same aihwkit device-level noise model over 50 Monte-Carlo programming runs; each configuration reports the mean, standard deviation, worst-case, and 5th percentile accuracy, together with the ideal (noise-free) accuracy of the same digital model.

2.5 Analytical view of noise propagation

Consider an L -layer network with weight matrices $\mathbf{W}^{(l)}$ and activation $\phi(\cdot)$. For analytical tractability, we consider a simplified zero-mean independent perturbation model $\Delta W_{ij}^{(l)} \sim \mathcal{N}(0, \sigma_n^2)$. A perturbation injected at layer l propagates to the output through the downstream weight matrices and the local slope of the activation, so the output error admits a first-order scaling estimate of the form:

$$\|y - \tilde{y}\|_2 \lesssim \sigma_n \sum_{l=1}^L c_l \cdot \|\mathbf{h}^{(l-1)}\|_2, \quad (4)$$

$$c_l = \prod_{k=l}^L \left(\max_i |\phi'(\mathbf{W}^{(k)} \mathbf{h}^{(k-1)})| \cdot \|\mathbf{W}^{(k)}\| \right),$$

where c_l is the per-layer noise-accumulation factor given by the product of downstream weight norms and local activation slopes. This first-order estimate provides an analytical interpretation of on-chip robustness: the error contributed by any layer is amplified by the accumulated factor c_l , which accumulates multiplicatively across downstream layers. PQAT-FT can indirectly reduce the effective accumulation factor, by concentrating weights toward the low magnitude regime and leaving unused conductance headroom, it reduces the effective accumulation factor c_l and suppresses the multiplicative cascade, achieving physical noise clamping at the hardware level (Fig. 1c). This provides an analytical motivation for the benefit observed across the four activations tested.

3. RESULTS

3.1 Simulation noise robustness

Before fabrication, we isolate the interaction among quantization precision, training strategy, and activation under controlled perturbations through behavioral simulation (50-trial Monte Carlo). Fig. 2a and Fig. 2b overlay prediction trajectories: the QAT 4-bit baselines suffer catastrophic collapse with peak-to-peak deviations exceeding $\pm 50\%$ under RRAM noise (a), whereas the PQAT-FT 4-bit variants recover a robust trace with tight trajectory bundling (b), for both ReLU and Tanh.

Fig. 2c and Fig. 2d contrast the Tanh raw weight distributions under the 4-bit QAT baseline and the PQAT-FT variant. The 4-bit QAT baseline (c) spreads raw weights toward the high magnitude regime ($|w| \gtrsim 0.8$), which can produce larger pre-activation magnitudes and thereby increase the likelihood of saturation, where local gradients vanish; the PQAT-FT variant (d) reshapes the distribution toward zero through sparse regularization, leaving the high magnitude regime vacant to serve as a noise margin.

Fig. 2e and Fig. 2f trace the bit width scaling of simulated RRAM RMSE for ReLU (e) and Tanh (f), comparing SW, QAT, and PQAT-FT across 2/4/6/8-bit. PQAT-FT stays below its QAT baseline at every precision. The QAT curves are markedly non-monotonic, Tanh QAT climbs from 35.3 (2-bit) to 65.5 (4-bit) and 96.9 (8-bit), whereas the PQAT-FT curves remain flat and low across the whole range (28.0, 22.9, 25.4, 26.0 for Tanh).

Table 1 summarizes the full simulation matrix across four activations and 2/4/6/8-bit precision. PQAT-FT improves every activation at 4 bits relative to its QAT baseline, with

Tanh showing the largest relative RMSE reduction (65.5 $\rightarrow$ 22.9) and reaching a low simulated RMSE (22.9 $\pm$ 1.3) even below its 6-bit counterpart (25.4 $\pm$ 1.2), anticipating the counter-intuitive result later confirmed on chip.

Fig. 2g quantifies the three-sigma error bound ($\mu + 3\sigma$) across the four activations at 4-bit. For every activation the PQAT-FT bound lies well below both the SW and QAT bounds: Tanh tightens from 127.7 (SW) and 76.5 (QAT) to 26.9, ReLU from 52.7/43.0 to 26.8, Leaky ReLU from 45.8/47.2 to 29.6, and Sigmoid from 34.3/33.1 to 23.3, confirming that the improvement extends to the tail of the noise distribution rather than the mean alone.

Fig. 2h reports the relative recovery of PQAT-FT over direct floating-point mapping, $(RMSE_{SW} - RMSE_{FT})/RMSE_{SW}$, across activations and bit widths. The recovery is positive in every cell and largest for Tanh (up to 79% at 4-bit); the unbounded activations still recover 40–49% (ReLU) and 25–33% (Leaky ReLU), and Sigmoid 20–37%. This indicates that the noise-margin mechanism transfers across activation families rather than depending on boundedness.

3.2 Hardware demonstration and on-chip validation

Fig. 3a displays the TiN/TaOx/HfOx/TiN 8K RRAM CIM test chip mounted on its custom testing PCB board, onto which the three-layer MLP predictor ($7 \rightarrow 20 \rightarrow 20 \rightarrow 1$) is mapped in the single device reference-column scheme.

Fig. 3b presents the layer-1 conductance mapping for Tanh PQAT-FT 4-bit: target conductance, actual programmed values measured on-chip, and absolute programming error $|\Delta G|$. The root-mean-square programming error is 5.3 μS , i.e., a mean cell-wise relative error $|\Delta G|/G_{\text{target}}$ of 2.8%, characterizing the programming accuracy and the physical conductance write noise under the adopted analog mapping. The largest localized deviations remain below 27 μS , appearing as bright spots in the error heatmap. These localized programming deviations provide a physical indication of the non-idealities that can contribute to signal degradation analyzed in Section 2.5.

Fig. 3c overlays the 132-month hardware inference trajectories on the ground-truth sunspot time series. The Tanh PQAT-FT 4-bit prediction tracks the ground truth across the full horizon with a low on-chip RMSE (34.6 $\pm$ 5.5), deviating primarily at cycle extrema where the dynamics are most nonlinear. By contrast, the Tanh QAT 4-bit baseline degrades substantially (96.5 $\pm$ 12.4), the largest relative improvement among the four activations, confirming that conventional QAT, though effective in the ideal digital domain, does not prepare the network for device-level programming noise.



Figure 2 – Behavioral simulation under the aihwkit device-level noise model (50-trial Monte Carlo at inference with state-independent programming noise; additive Gaussian $\sigma_n = 0.1$ is injected only during noise-aware fine-tuning). (a)–(b) First 80 months of the 132-month test horizon. (a) QAT 4-bit baselines under RRAM noise: catastrophic prediction collapse with peak-to-peak deviations exceeding $\pm 50\%$ for both ReLU and Tanh. (b) PQAT-FT 4-bit: robust trace recovery with tight trajectory bundling for both activations. (c)–(d) Tanh raw weight histograms: QAT 4-bit (c) spreads raw weights across a wide range, which can enlarge pre-activation magnitudes and thus increase the likelihood of saturation, whereas PQAT-FT 4-bit (d) concentrates them near zero through sparse regularization, leaving a physical conductance margin in the high headroom region. (e)–(f) Bit width scaling (2/4/6/8-bit) of RRAM RMSE for ReLU (e) and Tanh (f), comparing SW, QAT, and PQAT-FT. (g) Three-sigma error bound ($\mu + 3\sigma$) across four activations at 4-bit. (h) Relative recovery of PQAT-FT over direct floating-point mapping, $(RMSE_{SW} - RMSE_{FT})/RMSE_{SW}$, across activations and bit widths. The $\mu + 3\sigma$ and relative-recovery percentages are derived from unrounded Monte-Carlo statistics.

Table 1 – Simulated RRAM-inference RMSE across four activations and bit widths (50-trial Monte Carlo; mean $\pm$ standard deviation).

Activation	SW	QAT-2	QAT-4	QAT-6	QAT-8	FT-2	FT-4	FT-6	FT-8
Sigmoid	31.9 $\pm$ 0.8	51.0 $\pm$ 1.1	30.8 $\pm$ 0.8	32.7 $\pm$ 1.1	26.0 $\pm$ 1.3	20.1 $\pm$ 0.9	20.5 $\pm$ 1.0	20.5 $\pm$ 0.9	25.5 $\pm$ 0.9
ReLU	44.2 $\pm$ 2.8	57.2 $\pm$ 2.0	34.3 $\pm$ 2.9	41.5 $\pm$ 3.0	43.8 $\pm$ 2.7	26.4 $\pm$ 1.9	22.7 $\pm$ 1.4	25.0 $\pm$ 1.2	23.2 $\pm$ 1.5
Tanh	110.0 $\pm$ 5.9	35.3 $\pm$ 1.4	65.5 $\pm$ 3.7	91.3 $\pm$ 5.9	96.9 $\pm$ 6.0	28.0 $\pm$ 1.3	22.9 $\pm$ 1.3	25.4 $\pm$ 1.2	26.0 $\pm$ 1.7
Leaky ReLU	37.9 $\pm$ 2.6	54.8 $\pm$ 1.7	37.5 $\pm$ 3.3	43.1 $\pm$ 3.3	46.9 $\pm$ 3.5	27.7 $\pm$ 2.1	25.3 $\pm$ 1.5	25.8 $\pm$ 1.8	28.5 $\pm$ 2.2

Note: SW: standard FP32 full-precision training; QAT: quantization-aware training; FT: Pretrained QAT with noise-aware Fine-Tuning (PQAT-FT). Numbers in the QAT/FT columns report the 2/4/6/8-bit variants. All RMSE values are in physical sunspot units, rounded to one decimal place.

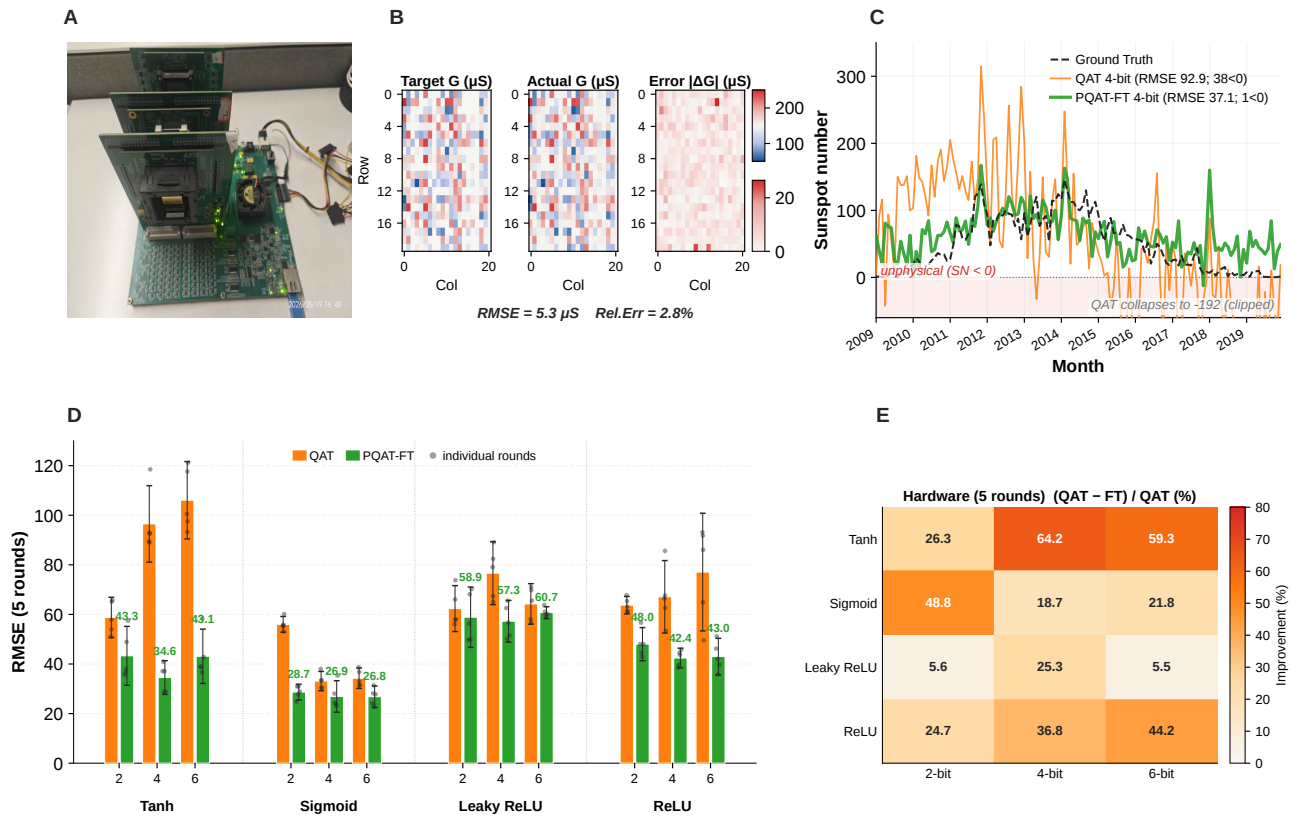


Figure 3 – On-chip measurements on the fabricated 8K RRAM CIM test chip. (a) Chip photo on its custom testing PCB board. (b) Layer-1 conductance heatmaps for Tanh PQAT-FT 4-bit: target conductance, actual programmed values, and absolute programming error $|\Delta G|$. (c) 132-month hardware inference waveforms for Tanh QAT 4-bit versus PQAT-FT 4-bit (a representative programming round is shown; Table 2 reports the five-round mean $\pm$ SD); the QAT trace is clipped where it diverges beyond the plotted range. (d) On-chip RMSE over five independent hardware rounds (bars: mean $\pm$ 95% confidence interval; dots: individual rounds; green values: PQAT-FT five-round means) for four activations at 2/4/6-bit. (e) Hardware improvement heatmap, $(QAT-FT)/QAT$ (%), across activations and bit widths. Improvement percentages are calculated from unrounded five-round statistics.

To mitigate single measurement bias and ensure statistical repeatability, each configuration is programmed and measured over five independent rounds; Fig. 3d reports mean $\pm$ 95%-confidence-interval bars with individual round markers. Table 2 reports the full hardware matrix as mean $\pm$ standard deviation across the five rounds. Across all four activation functions, PQAT-FT improves on QAT at every tested bit width; at 4 bits the reductions are 64.2% (Tanh), 36.8% (ReLU), 18.7% (Sigmoid), and 25.3% (Leaky ReLU). Fig. 3e summarizes the hardware improvement heatmap, $(QAT-FT)/QAT$, across activations and bit widths: the relative gain is positive in every cell and largest at 4-bit Tanh, confirming that the benefit is not confined to bounded activations.

3.3 On-chip comparative results

A counter-intuitive result emerges from Table 2: the 4-bit Tanh variant (34.6 ± 5.5) outperforms its 6-bit counterpart (43.1 ± 8.8) by 19.7%, inverting the conventional accuracy–precision trade-off. We examine this further in Section 4.

3.4 Weight distribution analysis

The Tanh QAT 4-bit baseline (RMSE 96.5) performs poorly despite sharing Tanh’s bounded activation with PQAT-FT. Without Stage-3 noise-aware fine-tuning, the weights disperse across the full range into the low headroom region ($|w| \geq 0.8$), mapping physical conductances dangerously close to the $[50, 250] \mu\text{S}$ operational boundaries where device non-linearities and drift are pronounced. In contrast, the PQAT-FT variant concentrates weights predominantly within the high headroom region ($|w| < 0.8$) through the combined effect of L1 regularization and noise injection. This confines the mapped conductances roughly within $[70, 230] \mu\text{S}$, deliberately carving out an unused physical conductance headroom of approximately $20 \mu\text{S}$ on either side. This reserved margin directly absorbs programming variations and physical fluctuations before they saturate the conductance state or perturb the subsequent layer currents.

Table 2 – On-chip RMSE across four activations and bit widths on the fabricated 8K RRAM CIM chip (five independent programming rounds; mean $\pm$ standard deviation).

Activation	SW	QAT-2	QAT-4	QAT-6	FT-2	FT-4	FT-6
Tanh	128.4 $\pm$ 24.9	58.8 $\pm$ 6.6	96.5 $\pm$ 12.4	106.1 $\pm$ 12.6	43.3 $\pm$ 9.6	34.6 $\pm$ 5.5	43.1 $\pm$ 8.8
Sigmoid	35.4 $\pm$ 3.5	56.0 $\pm$ 2.6	33.1 $\pm$ 3.1	34.3 $\pm$ 3.3	28.7 $\pm$ 2.6	26.9 $\pm$ 5.1	26.8 $\pm$ 3.5
ReLU	97.2 $\pm$ 7.5	63.7 $\pm$ 2.8	67.1 $\pm$ 11.8	77.1 $\pm$ 19.1	48.0 $\pm$ 5.4	42.4 $\pm$ 3.2	43.0 $\pm$ 5.9
Leaky ReLU	80.3 $\pm$ 22.1	62.4 $\pm$ 7.5	76.6 $\pm$ 10.2	64.2 $\pm$ 6.6	58.9 $\pm$ 9.8	57.3 $\pm$ 6.8	60.7 $\pm$ 1.9

Note: SW: standard FP32 full-precision training; QAT: quantization-aware training; FT: Pre-trained QAT with noise-aware Fine-Tuning (PQAT-FT). Each cell reports the RMSE mean $\pm$ standard deviation across five independent hardware programming rounds. All RMSE values are in physical sunspot units, rounded to one decimal place.

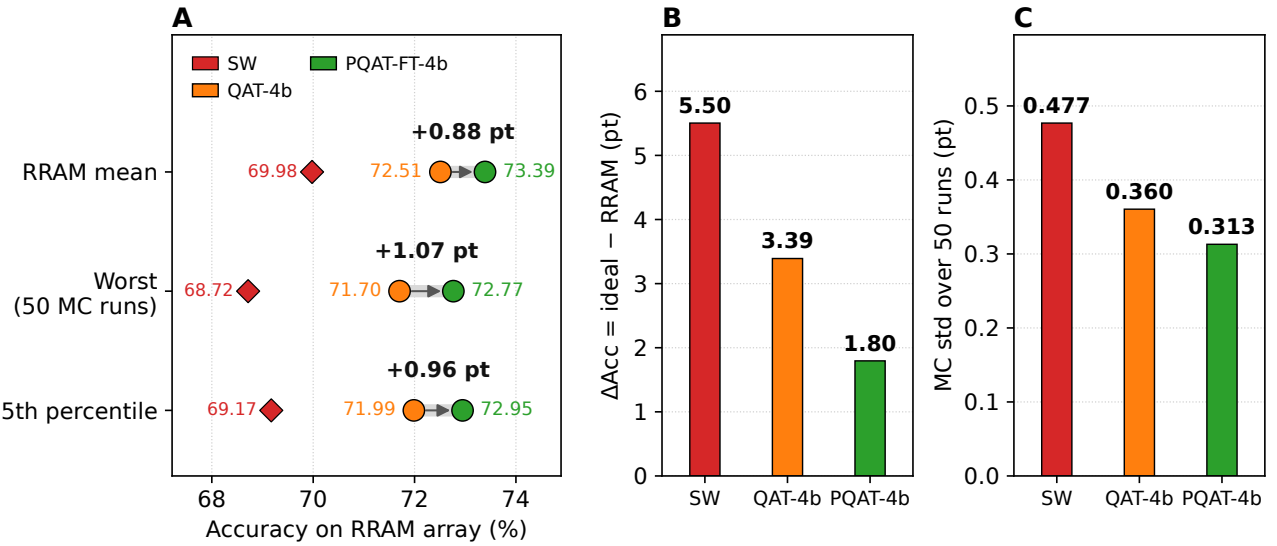


Figure 4 – Cross-task evaluation of PQAT-FT on the 2D EuroSAT remote-sensing benchmark (Leaky ReLU, 4-bit) under simulated RRAM noise (aihwkit behavioral model, not an on-chip measurement). (a) Accuracy on the simulated RRAM array, showing mean, worst-case, and 5th percentile metrics over the 50 Monte-Carlo evaluations. (b) Accuracy degradation relative to ideal software inference ($\Delta\text{Acc} = \text{ideal} - \text{RRAM}$, lower is better). (c) Monte-Carlo inference variability quantified by the standard deviation over the 50 Monte-Carlo evaluations. A representative run is shown, a single random initialization drives training and the 50-run Monte-Carlo evaluation, so the floating-point baseline (SW), QAT, and PQAT-FT share an identical hardware-noise stream; all metrics are computed on the held-out test split. Derived differences (percentage point gaps and improvement percentages) are calculated from unrounded data.

3.5 Cross-task transfer to a 2D benchmark

To probe the boundary of the mechanism beyond 1D forecasting, we transfer PQAT-FT to the 2D EuroSAT benchmark whose data and protocol are specified in Section 2.4. Fig. 4 reports the classification accuracy of the test split under simulated RRAM inference.

The 2D classification performance is evaluated over 50 Monte-Carlo inference trials using the aihwkit behavioral noise model. As shown in Fig. 4a, PQAT-FT raises the worst-case accuracy by 1.07 percentage points (71.70% $\rightarrow$ 72.77%) and the 5th percentile accuracy by 0.96 percentage points (71.99% $\rightarrow$ 72.95%) over QAT, despite a modest mean accuracy gain (72.51% $\rightarrow$ 73.39%). In terms of hardware penalty (Fig. 4b), PQAT-FT suppresses the accuracy degradation ($\Delta\text{Acc} = \text{ideal} - \text{RRAM}$) from 3.39 points in QAT down to 1.80 points; this reduction is not an artifact of a weaker software baseline, since the ideal noise-free accuracy differs by only 0.72 points (75.90% for QAT versus 75.19% for PQAT-FT)

while the RRAM accuracy itself is higher by 0.88 points. Furthermore, Fig. 4c reveals that PQAT-FT reduces the Monte-Carlo inference standard deviation by 13% (0.360 $\rightarrow$ 0.313), providing evidence that similar robustness-oriented behavior extends to the 2D vision task. Across three independent training runs, variance reduction was observed in all three runs, whereas lower-bound improvement was observed in two of the three runs.

The narrowed mean gain is expected and does not contradict the 1D result: classification decisions depend only on the relative ordering of class logits, so weight noise must be large enough to cross the decision boundary before it changes a prediction, whereas 1D regression accumulates every perturbation directly into the RMSE; moreover, the cross-entropy objective already encourages a decision margin, leaving less room for the noise-margin mechanism to improve the mean. The 2D task therefore calls for a different optimization emphasis than the 1D forecasting setting, and the narrowed mean gain bounds the generality of the 1D result.

3.6 Hardware efficiency context

To situate the energy benefit of RRAM CIM on a common footing, Table 3 compares this work's energy efficiency, 0.61 TOPS/W, i.e., 1.65 pJ per operation (pJ/op) over a 64×128 array including peripheral circuits, against three representative digital accelerators (NVDLA, Eyeriss, and an NVIDIA A100 GPU). Because the baselines are fabricated in far more advanced nodes, all platforms are normalized to a common 180 nm reference using the CMOS logic-gate (FO4 inverter) energy scaling factors of Stillmaker and Baas. After this node normalization, this work exceeds the digital baselines by $17\times$ (NVDLA), $58\times$ (Eyeriss), and $96\times$ (A100), despite its three-to-seven-generation node disadvantage; these figures are a first-order node comparison, not a fabricated measurement of the digital baselines. Area and latency figures are not directly comparable across paradigms and are therefore omitted; the FO4 scaling factors capture logic-gate energy trends only and do not account for memory-hierarchy or architecture-level energy scaling. The 16-fold gap between the cell-level (103 fJ) and peripheral-inclusive (1.65 pJ) figures arises mainly from the ADC/DAC conversion and readout circuitry, which dominate the per-operation energy in mixed-signal CIM.

4. DISCUSSION

The on-chip measurements establish noise-aware fine-tuning as a cross-layer robustness primitive across the four activations and bit widths tested on this hardware platform, rather than a remedy specific to any one activation: PQAT-FT reduces the 4-bit RMSE relative to QAT for all four activations tested, by 64.2% (Tanh), 36.8% (ReLU), 18.7% (Sigmoid), and 25.3% (Leaky ReLU). A plausible explanation for this consistency is the mechanism shared across all cases: sparse regularization and noise injection concentrate the trained weights toward the low magnitude regime and deliberately leave the high magnitude conductance space vacant. This noise buffer absorbs device-level fluctuations before they can drive synaptic weights into the low headroom boundary regions where device non-linearities and drift are pronounced, thereby mitigating the layer-by-layer error amplification that our mathematical analysis identifies as an important degradation pathway. These results demonstrate that proactive weight distribution shaping and conductance headroom, rather than activation choice alone, contribute substantially to the observed robustness.

The counter-intuitive superiority of 4-bit over 6-bit Tanh (34.6 ± 5.5 versus 43.1 ± 8.8 , a 19.7% margin) is consistent across all five independent on-chip programming rounds, uncovering a physically driven, non-monotonic precision-scaling behavior. This phenomenon is consis-

tent with a physical Signal-to-Noise Ratio (SNR) mismatch between the nominal quantization spacing and the intrinsic device programming noise. On our fabricated RRAM platform, each cell stores one analog weight and the measured RMS programming error is approximately $5.3 \mu\text{S}$. In a 6-bit configuration (64 nominal target conductances corresponding to 6-bit weight quantization across the $200 \mu\text{S}$ window), the nominal conductance spacing between adjacent targets is merely $\sim 3.2 \mu\text{S}$. Because the measured programming error ($5.3 \mu\text{S}$) exceeds the nominal step size, the 6-bit nominal targets may experience substantial overlap, which could corrupt the stored synaptic weights. Conversely, the 4-bit configuration (16 nominal target conductances corresponding to 4-bit weight quantization) provides a much coarser conductance spacing of $\sim 13.3 \mu\text{S}$, which provides a larger nominal margin to tolerate the $5.3 \mu\text{S}$ programming fluctuations. By pairing the 4-bit grid with PQAT-FT to anchor weights within the high headroom linear conductance region, the system may provide a more favorable effective SNR, which may explain why the 4-bit network outperforms the nominally higher 6-bit precision in a noisy physical environment.

It is telling that Tanh still yields the largest relative improvement over its QAT baseline. Bounded activations are not the source of PQAT-FT's cross-activation consistency but they are its strongest synergist. A hard output ceiling complements the noise buffer in two ways: it limits activation magnitude and can reduce local perturbation sensitivity, and it prevents any single perturbed activation from imposing an arbitrarily large excursion on downstream layers. Crucially, this synergism is a property of the physical system that simulation only partially foreshadows: in simulation the 4-bit-over-6-bit inversion is subtle (22.9 ± 1.3 versus 25.4 ± 1.2), whereas on chip it opens into a larger 19.7% gap. Physical conductance programming, with its finite level resolution and localized faults, amplifies precisely the regime in which a bounded output combined with a widened margin matters most. The activation function is accordingly best understood not as the primary defense but as the component that, combined with weight redistribution, yields the largest relative improvement over its QAT baseline. Whether this depends on boundedness per se or on a particular function shape remains to be established beyond the four functions examined.

The 2D EuroSAT benchmark clarifies the boundary of this mechanism. The mean accuracy gain is modest: variance reduction is observed in all three independent runs, while lower-bound (worst-case and 5th percentile) improvement is observed in two of the three runs, indicating that the robustness mechanism transfers even where its magnitude narrows. This asymmetry is expected from the task structure: classification is inherently tolerant to weight noise because only perturbations that cross the decision boundary change a prediction, and

Table 3 – Hardware metric comparison between this work (180 nm RRAM CIM) and representative digital accelerators, normalized to a common 180 nm node using CMOS logic-gate energy scaling factors.

Metric	This work	NVDLA[23]	Eyeriss[24]	NVIDIA A100[25]
Technology	180 nm	16 nm	65 nm	7 nm
Compute paradigm	RRAM CIM	digital NPU	digital ASIC	GPU
Bit precision	4-bit (weight quantization)	INT8	16-bit	INT8
Measured/reported EE (TOPS/W)	0.61 [†]	5.4	0.166	1.56
Measured/reported E/op (pJ/op)	1.65 [†]	0.185	6.0	0.64
180-nm normalized EE (TOPS/W) [‡]	0.61	0.035	0.0104	0.0063
180-nm normalized E/op (pJ/op) [‡]	1.65	28.5	96	158

Note: [†]The 103 fJ cell-level contribution is derived from the measured electrical parameters (0.185 V, 20 ns read pulse, 150 μ S base conductance), whereas the peripheral-inclusive energy is 1.65 pJ per operation over a 64×128 (8 192-cell) array under the adopted full-operation energy accounting, in which each multiply-accumulate is counted as two operations (one multiply and one accumulate), i.e., 0.61 TOPS/W. [‡]Node normalization. Reported per-operation energy follows from the reciprocal relation E_{op} (pJ) = $1/EE$ (TOPS/W) ($1/5.4 \approx 0.185$, $1/0.166 \approx 6.0$, $1/1.56 \approx 0.64$). Each digital baseline is then rescaled to the 180 nm node by the CMOS logic-gate (FO4 single-inverter) energy ratio of Stillmaker and Baas[26] (16 nm $\times 154$, 65 nm $\times 16$, 7 nm $\times 248$; 180 nm reference 27.5 fJ), i.e., normalized E/op = native E/op $\times$ factor (equivalently normalized EE = native EE $\div$ factor). This normalization assumes logic-gate-dominated energy and is applied in the favorable direction only (more advanced node $\rightarrow$ 180 nm); it is a first-order node comparison, not a fabricated-180 nm measurement of the digital baselines. Digital figures are traced to published datasheets or peer-reviewed literature (a source-to-value mapping is provided to the reviewers).

the cross-entropy objective already encourages a decision margin, whereas 1D regression accumulates every perturbation into the RMSE, leaving a larger scope for mean-error improvement.

Several limitations bound this study. The network architecture is fixed at $7 \rightarrow 20 \rightarrow 20 \rightarrow 1$; how activation choice interacts with depth, width, or skip connections remains to be characterized. The 1D task is single-output regression on a specific dataset, and the 2D task is a ten-class image classification; the failure modes may differ in other regression or multi-output settings. On-chip validation is limited to a single 8K test chip at room temperature; endurance cycling, radiation hardness, and thermal stress, critical for space qualification, remain to be assessed. The four activations tested represent a limited sample; whether the observed pattern extends to other bounded functions (e.g., hard tanh, softsign) or to smooth activations with partial bounding (e.g., Swish, GELU) requires further study.

5. CONCLUSION

Measurements on a fabricated 8K RRAM CIM test chip indicate that noise-aware fine-tuning acts as a hardware noise margin for low bit RRAM CIM: PQAT-FT improves on-chip RMSE over QAT for all four activations tested, including the unbounded functions (ReLU and Leaky ReLU), with 4-bit reductions of 64.2% (Tanh), 36.8% (ReLU), 18.7% (Sigmoid), and 25.3% (Leaky ReLU). Under Tanh, the 4-bit configuration (34.6 ± 5.5) outperforms the 6-bit configuration (43.1 ± 8.8) on this platform, departing from the conventional accuracy-precision trade-off. A 2D EuroSAT transfer shows a similar robustness-oriented trend, reduced Monte-Carlo inference variance and improved lower-bound accuracy, while its mean gain narrows, bounding the generality of the 1D result.

These findings support the potential of noise-aware fine-tuning as a cross-layer robustness primitive for low bit RRAM CIM toward future space-borne edge systems; radiation qualification remains a subject of future work.

ACKNOWLEDGEMENT

Acknowledgement: This work was supported by the National Natural Science Foundation of China under Grant Nos. U22A6001, U23A20365, 61821091, 61888102, and 61825404.

REFERENCES

- [1] Gabriela Mystkowska, Matteo Monopoli, Pietro Nannipieri, Luca Zulberti, David Merodio Codinachs, and Luca Fanucci. "Hardware Platforms Enabling Edge AI for Space Applications: A Critical Review". In: *IEEE Access* 13.1 (2025), pp. 143939–143956. doi: [10.1109/ACCESS.2025.3596326](https://doi.org/10.1109/ACCESS.2025.3596326).
- [2] Linfang Wang, Weizeng Li, Zhidao Zhou, Junjie An, Wang Ye, Zhi Li, Hanghang Gao, Hongyang Hu, Jing Liu, Xiaoming Chen, Ling Li, Qi Liu, Mingoo Seok, Chunmeng Dou, and Ming Liu. "A near-threshold memristive computing-in-memory engine for edge intelligence". In: *Nature Communications* 16.1 (July 2025), p. 5897. issn: 2041-1723. doi: [10.1038/s41467-025-61025-4](https://doi.org/10.1038/s41467-025-61025-4).
- [3] Fernando Aguirre, Abu Sebastian, Manuel Le Gallo, Wen-hao Song, Tong Wang, J. Joshua Yang, Wei Lu, Meng-Fan Chang, Daniele Ielmini, Yuchao Yang, Adnan Mehonic, Anthony Kenyon, Marco A. Villena, Juan B. Roldán, Yiming Wu, Hsiao-Hsuan Hsu, Nagarajan Raghavan, Jordi Suñé, Enrique Miranda, Mario Lanza, et al. "Hardware implementation of memristor-based artificial neural networks". In: *Nature Communications* 15.1 (2024). doi: [10.1038/s41467-024-45670-9](https://doi.org/10.1038/s41467-024-45670-9).
- [4] Hao Yang, Pingbing Zuo, Kun Zhang, Zhenning Shen, Zhengyang Zou, and Xueshang Feng. "Forecasting long-term sunspot numbers using the LSTM-WGAN model". In: *Frontiers in Astronomy and Space Sciences* 12 (2025). issn: 2296-987X. doi: [10.3389/fspas.2025.1541299](https://doi.org/10.3389/fspas.2025.1541299).
- [5] Tingrui Ren, Xvpeng Han, Kexin Yao, Huimeng Guo, Liang Wang, Yuanfu Zhao, and Bi Wang. "Radiation Effects Modeling and Analysis in RRAM-Based In-Memory Computing Systems".

- In: *2025 6th International Conference on Radiation Effects of Electronic Devices (ICREED)*. 2025, pp. 1–4. doi: [10.1109/ICREED65908.2025.11036223](https://doi.org/10.1109/ICREED65908.2025.11036223).
- [6] Debayan Mahalanabis, Hugh J. Barnaby, Michael N. Kozicki, Vineeth Bharadwaj, and Saba Rajabi. “Investigation of Single Event Induced Soft Errors in Programmable Metallization Cell Memory”. In: *IEEE Transactions on Nuclear Science* 61.6 (Dec. 2014), pp. 3557–3563. issn: 0018-9499, 1558-1578. doi: [10.1109/TNS.2014.2358235](https://doi.org/10.1109/TNS.2014.2358235).
 - [7] Seungkyu Choi, Jaekang Shin, and Lee-Sup Kim. “A Deep Neural Network Training Architecture With Inference-Aware Heterogeneous Data-Type”. In: *IEEE Transactions on Computers* 71.5 (2022), pp. 1216–1229. doi: [10.1109/TC.2021.3078316](https://doi.org/10.1109/TC.2021.3078316).
 - [8] Daniele Ielmini and H.-S. Philip Wong. “In-memory computing with resistive switching devices”. In: *Nature Electronics* 1.6 (June 2018), pp. 333–343. issn: 2520-1131. doi: [10.1038/s41928-018-0092-2](https://doi.org/10.1038/s41928-018-0092-2).
 - [9] M. Alayan, M. Bagatin, S. Gerardin, A. Paccagnella, L. Larcher, E. Vianello, E. Nowak, B. De Salvo, and L. Perniola. “Experimental and Simulation Studies of the Effects of Heavy-Ion Irradiation on HfO₂-Based RRAM Cells”. In: *IEEE Transactions on Nuclear Science* 64.8 (2017), pp. 2038–2045. doi: [10.1109/TNS.2017.2721980](https://doi.org/10.1109/TNS.2017.2721980).
 - [10] Yongxiang Li, Shiqing Wang, and Zhong Sun. “Constraints on Bit Precision and Row Parallelism for Reliable Computing-in-Memory”. In: *Electronics* 14.13 (2025). issn: 2079-9292. doi: [10.3390/electronics14132532](https://doi.org/10.3390/electronics14132532).
 - [11] Mingyi Rao, Hao Tang, Jiangbin Wu, Wenhao Song, Minxian Zhang, Wenbo Yin, Ye Zhuo, Fatemeh Kiani, Benjamin Chen, Xiangpeng Jiang, Heqian Liu, Hong-Yu Chen, Rivu Midya, Fan Ye, Hao Jiang, Zhongrui Wang, Mingche Wu, Miao Hu, Huaqiang Wang, J. Joshua Yang, et al. “Thousands of conductance levels in memristors integrated on CMOS”. In: *Nature* 615.7954 (2023), pp. 823–829. doi: [10.1038/s41586-023-05759-5](https://doi.org/10.1038/s41586-023-05759-5).
 - [12] Tuo Shi, Lili Gao, Yang Tian, Shuangzhu Tang, Jinchang Liu, Yiqi Li, Ruixi Zhou, Shiyu Cui, Hui Zhang, Yu Li, Zuheng Wu, Xumeng Zhang, Taihao Li, Xiaobing Yan, and Qi Liu. “Memristor-based feature learning for pattern classification”. In: *Nature Communications* 16.1 (Jan. 2025), p. 913. issn: 2041-1723. doi: [10.1038/s41467-025-56286-y](https://doi.org/10.1038/s41467-025-56286-y).
 - [13] Tuo Shi, Hui Zhang, Shiyu Cui, Jinchang Liu, Zixi Gu, Zhanfeng Wang, Xiaobing Yan, and Qi Liu. “Stochastic neuro-fuzzy system implemented in memristor crossbar arrays”. In: *Science Advances* 10.12 (Mar. 2024), eadl3135. issn: 2375-2548. doi: [10.1126/sciadv.adl3135](https://doi.org/10.1126/sciadv.adl3135).
 - [14] Kyuree Kim, Min Suk Song, Hwiho Hwang, Sungmin Hwang, and Hyungjin Kim. “A comprehensive review of advanced trends: from artificial synapses to neuromorphic systems with consideration of non-ideal effects”. In: *Frontiers in Neuroscience* 18 (Apr. 2024), p. 1279708. issn: 1662-453X. doi: [10.3389/fnins.2024.1279708](https://doi.org/10.3389/fnins.2024.1279708).
 - [15] Lili Gao, Qingying Ren, Jiawei Sun, Su-Ting Han, and Ye Zhou. “Memristor modeling: challenges in theories, simulations, and device variability”. In: *Journal of Materials Chemistry C* 9.47 (2021), pp. 16859–16884. issn: 2050-7526, 2050-7534. doi: [10.1039/D1TC04201G](https://doi.org/10.1039/D1TC04201G).
 - [16] Malte J. Rasch, Charles Mackin, Manuel Le Gallo, An Chen, Andrea Fasoli, Frédéric Odermatt, Ning Li, S. R. Nandakumar, Pritish Narayanan, Hsin-yu Tsai, Geoffrey W. Burr, Abu Sebastian, and Vijay Narayanan. “Hardware-aware training for large-scale and diverse deep learning inference workloads using in-memory computing-based accelerators”. In: *Nature Communications* 14.1 (Aug. 2023), p. 5282. issn: 2041-1723. doi: [10.1038/s41467-023-40770-4](https://doi.org/10.1038/s41467-023-40770-4).
 - [17] Jinchang Liu, Jian Lu, Shuangzhu Tang, Ruixi Zhou, Huiqin Ma, Bo Lyu, Yang Tian, Tuo Shi, and Qi Liu. “Error-aware probabilistic training for memristive neural networks”. In: *Nature Communications* 16.1 (Dec. 2025), p. 11494. issn: 2041-1723. doi: [10.1038/s41467-025-66240-7](https://doi.org/10.1038/s41467-025-66240-7).
 - [18] Tuo Shi, Lili Gao, Ruixi Zhou, Yang Tian, Pei Chen, Yanting Ding, Shuangzhu Tang, Huiqin Ma, Jian Lu, Hui Zhang, Zhanfeng Wang, Bo Lyu, Xumeng Zhang, Xiaobing Yan, and Qi Liu. “Fully memristive spiking neural network for energy-efficient graph learning”. In: *Science Advances* 11.19 (2025), eadv2312. doi: [10.1126/sciadv.adv2312](https://doi.org/10.1126/sciadv.adv2312).
 - [19] Suhong Moon, Kwanghyun Shin, and Dongsuk Jeon. “Enhancing Reliability of Analog Neural Network Processors”. In: *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 27.6 (2019), pp. 1455–1459. doi: [10.1109/TVLSI.2019.2893256](https://doi.org/10.1109/TVLSI.2019.2893256).
 - [20] Olga Krestinskaya, Mohammed E. Fouda, Ahmed Eltawil, and Khaled N. Salama. *Joint Hardware-Workload Co-Optimization for In-Memory Computing Accelerators*. 2026. arXiv: [2603.03880](https://arxiv.org/abs/2603.03880) [cs.AR].
 - [21] Meiran Zhao, Bin Gao, Jianshi Tang, He Qian, and Huaqiang Wu. “Reliability of analog resistive switching memory for neuromorphic computing”. In: *Applied Physics Reviews* 7.1 (Mar. 2020), p. 011301. issn: 1931-9401. doi: [10.1063/1.5124915](https://doi.org/10.1063/1.5124915).
 - [22] Malte J. Rasch, Diego Moreda, Tayfun Gokmen, Manuel Le Gallo, Fabio Carta, Cindy Goldberg, Kaoutar El Maghraoui, Abu Sebastian, and Vijay Narayanan. “A flexible and fast PyTorch toolkit for simulating training and inference on analog crossbar arrays”. In: *2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS)*. 2021, pp. 1–4. doi: [10.1109/AICAS51828.2021.9458497](https://doi.org/10.1109/AICAS51828.2021.9458497).
 - [23] NVIDIA Corporation. *NVIDIA Deep Learning Accelerator (NVDLA)*. NVIDIA Open-Source Hardware, HotChips 2018. 5.4 TOPS/W (ResNet-50, small configuration). 2018.
 - [24] Yu-Hsin Chen, Tushar Krishna, Joel S. Emer, and Vivienne Sze. “Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks”. In: *IEEE Journal of Solid-State Circuits* 52.1 (2017), pp. 146–158.
 - [25] NVIDIA Corporation. *NVIDIA A100 Tensor Core GPU Architecture*. NVIDIA Whitepaper. 624 TOPS INT8 (sparse) at 400 W. 2020.
 - [26] Aaron Stillmaker and Bevan Baas. “Scaling equations for the accurate prediction of CMOS device performance from 180 nm to 7 nm”. In: *Integration, the VLSI Journal* 58 (2017), pp. 74–81. doi: [10.1016/j.vlsi.2017.03.002](https://doi.org/10.1016/j.vlsi.2017.03.002).

AUTHORS



YANG TIAN is an assistant researcher at the Research Center for Computing in Intelligent Manufacturing, Zhejiang Lab. He received his Ph.D. degree in materials physics and chemistry from Zhejiang University in 2021. His research interests focus on intelligent computing technologies based on emerging memristive material, device physics, and system-level applications.



DUNAN HU is currently a Ph.D. candidate at State Key Laboratory of Silicon and Advanced Semiconductor Materials, School of Materials Science and Engineering, Zhejiang University, China. She received her B.S. degree from China University of Mining and Technol-

ogy in 2023. Her research interests focus on artificial synaptic devices.



LINGZI ZHANG is a hardware engineer at Zhejiang Lab. With over five years of professional experience, she specializes in complex SiP design, advanced hardware system architecture for both analog and digital boards, and seamless FPGA/software co-development.



LILI GAO is a senior research specialist at Zhejiang Lab. Her research primarily focuses on RRAM-based hardware-software co-design and computing optimization.



KEQUAN XIA is currently a researcher at Imperial College London. He received his Ph.D. degree from Zhejiang University in 2022. His research interests include Triboelectric Nano-Generators (TENGs), self-powered sensors, and triboelectric materials. He has authored or co-authored 46 technical papers.



TUO SHI is a professor at Zhejiang University. His current research interest focuses on neuromorphic computing based on memristors, including materials, devices, and applications.



JIANGUO LU is an associate professor at State Key Laboratory of Silicon and Advanced Semiconductor Materials, School of Materials Science and Engineering, Zhejiang University, China. His research interests mainly lie in the area of semiconducting films and optoelectronic devices, electrochemical energy storage and integrated systems, and nanoscale materials and smart coating. He was awarded the Distinguished Young Scholar of Zhejiang Province. He serves as the Editor of Chinese Chemical Letters, Tungsten, and Materials. He published more than 300 papers with citation times over 12000 and H-index of 60. He is one of the most highly cited Chinese researchers and the world's top 2% scientists.